Feature

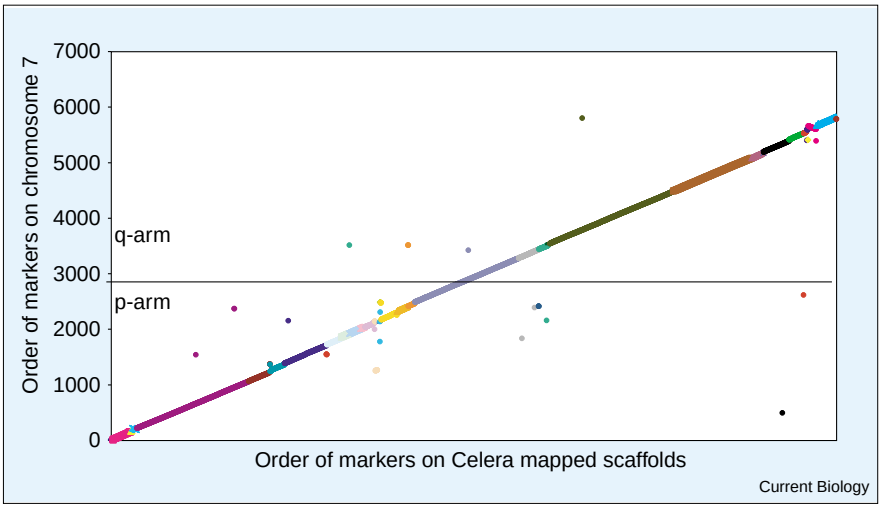
Discovery of the human genome sequence in the public and private databases

Genomes: Much heat has been generated in discussions about the key human genome sequence databases, generated by the Human Genome Project and Celera, and what specific features each offers genome researchers. Stephen W. Scherer and Joseph Cheung, who are intense users of both, offer a personal assessment of the developing contents.

Deservedly, there has been much celebration over the publication of two draft versions of the human genome sequence. There have also been other recent assemblies of the sequence, producing more complete coverage and reliable DNA sequence annotation. However, to date, a finished reference sequence of the human genome does not exist. Furthermore, only a fraction of the genes and other important biological features of chromosomes have been characterized. The goal of this piece is to share our experiences with other scientists contemplating if and how they might benefit from subscribing to the Celera DNA sequence database.

Our observations are based on having access to the Celera Discovery System through an 'academic' subscription for the past year. We are also intense users (and contributors) of data in the Human Genome Project (HGP) databases. Many of our experiences are based on gene mapping and sequencing studies of human chromosome 7, but also through positional cloning studies in other regions of the genome. We are most often asked to comment subjectively on the following three datasets:

Figure 1



The order of 5343 chromosome 7 DNA markers present in the C4 scaffolds (each scaffold in a different color) was almost entirely consistent with the marker order established by hand-curated data from radiation and somatic cell hybrid, yeast and bacterial- artificial chromosome, and genetic mapping experiments. The 246 markers that did not fall into these larger scaffolds were all found in smaller ones or in the Celera fragment database. The 22 DNA markers that are not in the expected order tend to map to the centromere or to intrachromosomal duplications. Over 98% of known markers could be placed on the map.

Important DNA sequence sources

(i) The Celera version of the human genome published in February (called component-3 or C3; data at http://www.celera.com/) and their more recent component-4 (C4) assembly (by subscription since August 2001). The C3 assembly was derived from combining 14,808 Mb of Whole Genome Shotgun (WGS) Celera sequence with 4,405 Mb from the HGP. C4 builds on C3 using improved algorithms as well as additional Celera sequences and new HGP data as of December 2000;

(ii) The successive assemblies of the clone-based approach of the HGP from the February publication up until August 2001 (up-to-date statistics for the HGP sequence can be found at http://www.ebi.ac.uk/genomes/mot/). The best websites for accessing HGP data are listed in Table 1. HGP does not have Celera data in their assemblies;

(iii) The Celera mouse genome (available since June 2001), assembled solely using a WGS based on approximately 6X genome coverage with DNA from three different mouse strains.

Table 1

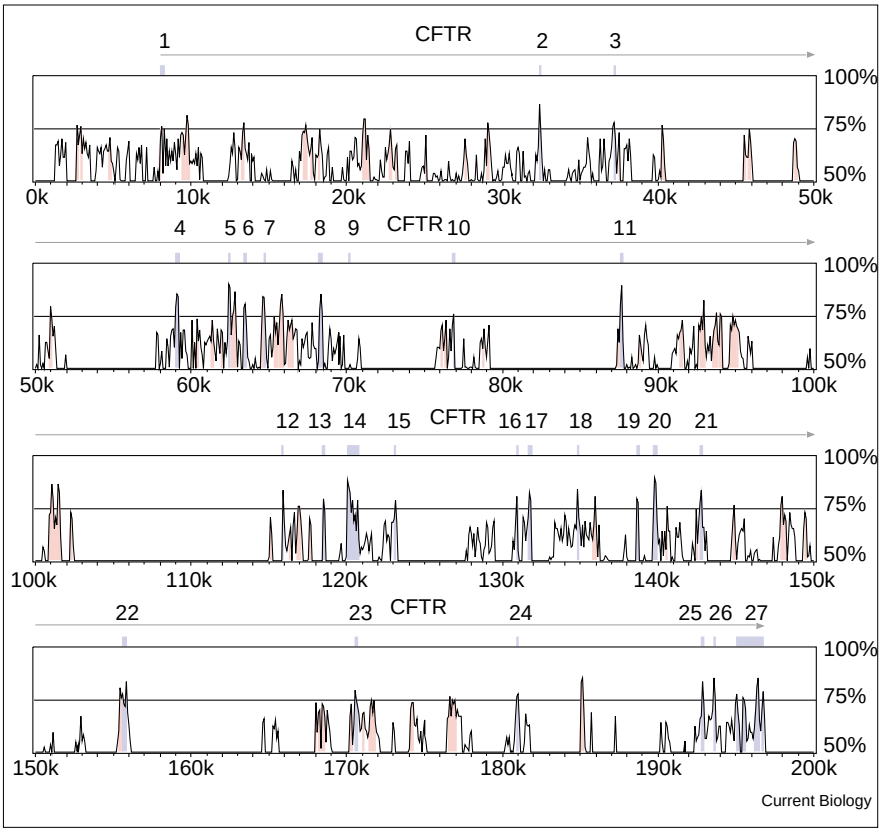
General characteristics of the Celera and HGP sequence databases.*

Category	Celera	Human genome project
Accessibility†		
To data	Good	Good to excellent
Via cytolocation	Very good (mirrors public data)	Very good
Via gene or marker	Good	Excellent
Via DNA sequence	Excellent	Good
Coverage		
Euchromatin	Outstanding	Good (~50% still in draft)
Pericentromeric	Good	Good
Large duplication	Not represented	Better than Celera
Accuracy		
Internal accuracy	Excellent	Excellent
Long-range order and orientation	Outstanding	Good, continues to improve
Gene annotation‡		
Known genes	Very good	Very good
New genes	Rudimentary	Rudimentary
Other strengths§		
	DNA sequence in fragment database often assists in gap filling	Ease of accessibility to data at multiple websites
	Long sequence scaffolds favor genome-wide comparison/annotation	Availability of clones to confirm or complete sequencing and mapping
	Availability of assembled mouse sequence to assist human annotation	Clone-based strategy essential for completion of difficult regions
Recommendations (wish list)		
	Be more dynamic incorporating latest public data	Increase resolution and accuracy of cytolocations
	Make clones available for sequencing of gap regions	Top up and finish human sequence
	Release human component 4 and mouse data on DVD to academic subscribers	Increase efforts to incorporate highly-curated data from community
	Sequence a third mammalian genome to assist comparative analyses	Sequence a third mammalian genome to assist comparative analyses

*Based on survey of 10 users of varying levels of sophistication; bioinformatics analysts (4), molecular biologists (3), medical geneticists (3). †While the Celera database is generally user friendly with excellent service support the limited number of portals per academic subscription can inhibit accessibility. Data retrieval can sometimes be slow. Navigating/searching HGP databases is more intuitive primarily since familiar nomenclature is used compared with the obscure identifiers often found in Celera. The favorite entry points to public DNA sequence data based on cytolocation, gene marker, and by DNA sequence itself are UCSC Golden Path (http://genome.ucsc.edu/) and the BAC resources (http://www.ncbi.nlm.nih.gov/genome/cyto/hbrc.shtml), Locus Link at NCBI (http://www.ncbi.nlm.nih.gov/LocusLink/) and again Golden Path, respectively. 'Ensembl' (http://www.ensembl.org/) was also a good entry point into the public data. Medical geneticists often use the Genome Database (http://www.gdb.org/) or GeneCards (http://bioinfo.weizmann.ac.il/cards/index.html).‡It is still premature to comment on the accuracy and completeness of the overall annotation of genes since many are based only on gene-prediction algorithms. Earlier versions of the Celera and HGP assemblies/annotation often missed contiguous gene family members but in both cases this continues to improve. For academic subscribers the supporting evidence for new Celera transcripts is not intuitively available to the end user. HGP is also more dynamic in updating new cDNA and gene data from the literature. §There are multitudes of DNA sequence analysis programs available in the public domain not mentioned (see http://searchlauncher.bcm.tmc.edu/). Celera has basic Blast search capabilities, GO ontology, Panther Ontology (proprietary), and Genome Browser which is an outstanding (proprietary) gene-model building tool (available to corporate but not academic subscribers). Celera's mouse genome assembly used the 129X1/SvJ, DBA/2J, and A/J strains and the HGP is sequencing C57BL6/J.

We have summarized our experiences using Celera compared to HGP information in Table 1. Both datasets, and the accompanying annotation, have strengths and weaknesses. While we constantly access both, if website hits alone were counted, the HGP would win out over Celera primarily because of ease of accessibility and increased number of entry points to the DNA sequence. In our group, more sophisticated analysts performing large-scale annotation experiments usually occupy our laboratory's single-portal access (per subscription) to Celera (the release of the C3 assembly on DVD has relieved some

Figure 2



Comparison of 200 kb of Celera human (C4) and mouse DNA sequence encompassing the cystic fibrosis (CFTR) gene on human chromosome 7 and mouse chromosome 6, respectively. Each window represents 50 kb of syntenic DNA sequence displayed using the program VISTA (http://www-gsd.lbl.gov/vista). Each of the 27 CFTR exons was present in the assembled mouse sequence. Blue shading represents exons and red highlights other highly conserved sequences.

of this pressure). Molecular biologists and medical geneticists almost always start by accessing the public databases to find out what can be found or what is missing, and then they check Celera. In some cases Celera's data is more complete and/or accurate than the HGP, in other cases it is not (Table 1).

For example, when annotating a chromosomal region for genes we most often use Celera sequence initially since it almost always represents longer continuous stretches of DNA sequence (scaffolds) than is currently found in the public database (Figure 1). This approach can lead to the identification of large (and sometimes small) genes that would have otherwise been fragmented or missing and, therefore, not detected using HGP data. For example, using Celera we have published manuscripts describing the CELSR2 (26 kb at 1p13–p21), RBM15 (8 kb at 1p13), c7orf10 (700 kb at 7p14), IMMP2L (860 kb at 7q31), RAY1/ST7 (220 kb at 7q31), CORTBP2 (170 kb at 7q31), and CASPR2 (2300 kb at 7q35) genes that, at the time, were not properly represented in HGP data. Our analysis of over 100 known full-length genes on chromosome 7 indicate they encompass an average of 50 kb of DNA (consistent with a chromosome 21 gene size of 57 kb). This suggests that the annotation of genes, in particular by the HGP, will become more accurate as the genome sequence moves from draft to finished form. As Hogenesch and colleagues have shown, however, the current Celera and Ensembl (HGP) sets of predicted genes are largely mutually exclusive, suggesting that even when a consensus genome sequence is achieved, the resulting gene maps will still vary greatly.

An example of where the HGP clone-based strategy outperforms the Celera WGS approach is in proper assembly of large nearly identical DNA segments that occur in more than one copy in the genome. Such duplications might account for up to 5% of human DNA. When duplications are >50 kb in size, in our experience, they are not represented in large C3 or C4 scaffolds (they are found in the Celera 'fragment' database). The same sequences may also be underrepresented or mistakenly assembled by the HGP.

However, we have found the HGP data usually to be more representative for these chromosomal regions with the added advantage of having access to a physical resource (the clone) for confirmatory analyses. For example, duplications involved in Williams–Beuren syndrome at 7q11.23 are not represented in Celera scaffolds, but they are better covered by the HGP. The same seems to be true for duplications flanking microdeletion and pericentromeric regions, as well as polymorphic genomic duplications such as those observed on chromosome 15 in panic disorder. As in the latter case, some discrepancies found in different versions of the genome may occur due to variation existing between the source(s) of DNA analyzed.

Importance of the mouse

The availability of the Celera mouse genome sequence has already

become an indispensable resource for interpreting the human genome. We have tested 952 human chromosome 7 genes and found 832 (87%) of the mouse orthologs to be accurately assembled into scaffolds assigned to 8 different murine chromosomes (six representing known syntenies and two requiring confirmatory mapping). The murine sequence has been instrumental in defining human gene structure (Figure 2), finding new genes, annotating regulatory regions, and of course in biological studies of the mouse.

In addition, since many of the problematic duplications in the human genome described earlier are relatively recent in origin (occurring after divergence of mouse and human), the mouse sequence can often serve as a ruler to refine the human sequence. The HGP is also sequencing the mouse genome using a combined WGS and clone-based strategy, but an assembled genome sequence has not yet been obtained.

Incremental gains

So, in the end, until someone completes a definitive version of the human genome comparable to that available for chromosome 21 and 22, but also with comprehensive annotation, the question "which is better" remains irrelevant. Any advantage the HGP or Celera might have over the other is incremental in nature.

Gains by the HGP are usually small but swift, while Celera's are massive but less dynamic. In fact, much of the discovery is fueled by having the ability to compare, contrast, and combine the different versions of the genome. For the past 12 months the availability of large amounts of human sequence at Celera not yet in the public databases, more than justified our investment. We anticipate the same accelerated rate of discovery over the next year by having access to an assembled mouse genome otherwise not available in the public domain.

Ultimately, as the absolute value of base pairs level out, the true measurement of value in these or any other databases will come from achieving a much higher level of DNA sequence, gene, and protein annotation, beyond what is now available.

Further reading

International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome sequence. *Nature* 2001, 409:860-921.

Venter JC, Adams MD, Meyers EW, Li PW., Mural PJ, Sutton GG, *et al.*: The sequence of the human genome. *Science* 2001, 291:1304-1361.

Green ED, Chakravarti A: The human genome sequence expedition: views from the "base camp". *Genome Res* 2001, 11:645-651.

Katsanis N, Worley KC, Lupski JR: An evaluation of the draft human genome sequence. *Nat Genet* 2001, 29:88-91.

Hogenesch JB, Ching KA, Batalov S, Su AI, Walker JR, Zhou Y, *et al.*: A comparison of the Celera and Ensemble predicted gene sets reveals little overlap in novel genes. *Cell* 2001, 106:413-415.

Eichler EE: Segmental duplications: what's missing, misassigned, and missassembled – and should we care? *Genome Res* 2001, 11:653-656.

Gratacos M, Nadal M, Martin-Santos R, Pujana MA, Gago J, Peral B, *et al.*: A polymorphic genomic duplication on human chromosome 15 is a susceptibility factor for panic and phobic disorders. *Cell* 2001, 106:367-379.

Stein L: Genome annotation: from sequence to biology. *Nat Rev Genet* 2001, 2:493-503.

Address: Genetics and Genomic Biology, The Hospital for Sick Children, Toronto, Canada.